mPLUG-Owl3: Towards Long Image-Sequence Understanding in Multi-Modal Large Language Models

Year: 2024 | *Citations: 208* | *Authors: Jiabo Ye, Haiyang Xu, Haowei Liu*

Abstract

Multi-modal Large Language Models (MLLMs) have demonstrated remarkable capabilities in executing instructions for a variety of single-image tasks. Despite this progress, significant challenges remain in modeling long image sequences. In this work, we introduce the versatile multi-modal large language model, mPLUG-Owl3, which enhances the capability for long image-sequence understanding in scenarios that incorporate retrieved image-text knowledge, interleaved image-text, and lengthy videos. Specifically, we propose novel hyper attention blocks to efficiently integrate vision and language into a common language-guided semantic space, thereby facilitating the processing of extended multi-image scenarios. Extensive experimental results suggest that mPLUG-Owl3 achieves state-of-the-art performance among models with a similar size on single-image, multi-image, and video benchmarks. Moreover, we propose a challenging long visual sequence evaluation named Distractor Resistance to assess the ability of models to maintain focus amidst distractions. Finally, with the proposed architecture, mPLUG-Owl3 demonstrates outstanding performance on ultra-long visual sequence inputs. We hope that mPLUG-Owl3 can contribute to the development of more efficient and powerful multimodal large language models.